Optimizing Multiscale SSIM for Compression via MLDS

Christophe Charrier, Kenneth Knoblauch, Laurence T. Maloney, Alan C. Bovik, Anush K. Moorthy

Optimizing Multi-Scale SSIM for Compression via MLDS

Christophe Charrier, *Member, IEEE,* Kenneth Knoblauch, Laurence T. Maloney,
Alan C. Bovik, *Fellow, IEEE* Anush K. Moorthy *Member, IEEE,*

Abstract

A crucial step in the assessment of an image compression method is the evaluation of the perceived quality of the compressed images. Typically, researchers ask observers to rate perceived image quality directly and use these rating measures, averaged across observers and images, to assess how image quality degrades with increasing compression. These ratings in turn are used to calibrate and compare image quality assessment algorithms intended to predict human perception of image degradation. There are several drawbacks to using such omnibus measures. First, the intepretation of the rating scale is subjective and may differ from one observer to the next. Second, it is easy to overlook compression artifacts that are only present in particular kinds of images.

In this paper, we use a recently developed method for assessing perceived image quality, Maximum Likelihood Difference Scaling (MLDS), and use it to assess the performance of a widely-used image quality assessment algorithm, MS-SSIM. MLDS allows us to quantify supra-threshold perceptual differences between pairs of images and to examine how perceived image quality, estimated through MLDS, changes as the compression rate is increased. We apply the method to a wide range of images and also analyze results for specific images. This approach circumvents the limitations inherent in the use of rating methods and allows us also to evaluate MS-SSIM for different classes of visual image. We show how the data collected by MLDS allows us to recalibrate MS-SSIM to improve its performance.

Index Terms

Image quality assessment performance, Difference scaling.

I. Introduction

Lossy image compression techniques such as JPEG2000 allow high compression rates, but only at the cost of perceived degradation in image quality. There is a considerable literature concerning how human observers perceive compression-induced degradation in images and how well several Image Quality Assessment (IQA) algorithms tend to predict human judgments of reduction in image quality as a function of compression.

C. Charrier is with Université de Caen-Basse Normandie, GREYC UMR CNRS 6072, Equipe Image, ENSICAEN, Caen, France.
K. Knoblauch is with INSERM, U846, Stem Cell and Brain Research Institute, Département Neurosciences Intégratives, Bron, France.
L.T. Maloney is with Department of Psychology, Center for Neural Science, New York University.
A.C. Bovik and A.K. Moorthy are with University of Texas at Austin, LIVE lab, Austin, TX.
This research is supported by the ANR project #ANR-08-SECU-007-04, and by the Intel and Cisco Inc under the VAWN program.

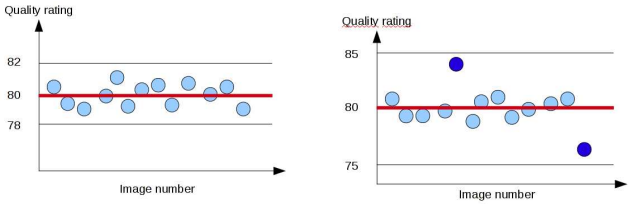
The most commonly employed means to assess human judgment of image quality is to ask human observers to rate image quality directly on a numerical scale. Human judgments are ordinarily expressed as the Mean Opinion Score (MOS) obtained from a sufficiently large set of human observer ratings relative to a normalized scale defined by the International Telecommunications Union (ITU) [**?**].

The typical summary of the agreement between rated subjective image quality and the ouput of an IQA algorithm is some measure of the correlation between the subjective ratings and the measured degree of distortion. Typical measures of correlation include 1) Pearson's linear correlation coefficient (CC) between MOS and algorithm score after nonlinear regression, 2) the root-mean-squared error (RMSE) between MOS and the algorithm score after nonlinear regression and 3) the Spearman rank order correlation coefficient (SROCC).

Examples of well-known IQA algorithms include DCtune [**?**], Picture Quality Scale (PQS) [**?**], Multi-Scale Structural SIMilarity (MS-SSIM) [**?**], Wavelet Structural Similarity (WSSI) [**?**], Visual Signal-to-Noise Ratio (VSNR) [**?**], and Visual Information Fidelity (VIF) [**?**] indices, to name a few. These indices compute relative quality scores between a reference image and a distorted version, often acheiving excellent correlations with MOS values. All those IQA indices have been designed using different frameworks. For example, MS-SSIM, WSSI and VIF were developed within a Natural Scene Statistics (NSS) framework or under assumptions about natural image structure. They are based on an assumption that distortion-free images occupy a small subspace of the space of all possible images. Image distortions can be interpreted as adding a distortion vector to distortion free images. DCtune and PQS were developed within a distortion-specific framework. They use distortion models based on a specific set of distortions (blockiness, blur, and so on) to predict quality scores. Any one of these algorithms can be judged better than a second if it correlates to a great extent with human MOS.

In [**?**], SHEIK *et al.* compared 10 recent IQA algorithms and determined which had particularly high levels of performance. They concluded that more can be done to reduce the gap between machine and human evaluation of image quality. In [**?**], SESHADRINATHAN and BOVIK studied the relationship between the structural similarity and VIF frameworks and older metrics, *i.e.* the MSE and HVS-based quality metrics. They concluded that SSIM and VIF are closely related to the older IQA metrics under certain natural scene modeling assumptions. This was, also, recently studied by HOR and ZIOU who defined a bijective relation between SSIM and PSNR yielding predictions of SSIM values from PSNR (and inversely) [**?**]. The global conclusion of all those comparison studies is that no IQA algorithm has been shown to definitively outperform all others for all possible degradations, although owing to the inclusion of both scene models and perceptual models, the MS-SSIM and VIF indices outperform many with statistical significance.

Consider two hypothetical IQA algorithms (say q_1 and q_2) that provide objective quality scores computed on a large database. Fig 1(a) and 1(b) illustrate non real samples of the obtained scores for each metric where outliers have been intentionally mentioned, for the purpose of our explanation. For each subfigure, let the score equal to 80 represents the ground truth score and the grey circles are the computed scores for each image using q_1 (Fig. 1(a)) and q_2 (Fig. 1(b)). Suppose the SROCC score is identical for the two metrics (say 0.96). This means that both IQA algorithms have the same global efficiency. Nevertheless, no information is provided about the "local" efficiency of



(a) distribution of the ratings obtained for the metric q_1.

(b) distribution of the ratings obtained for the metric q_2.

Fig. 1. Sample of the quality ratings obtained for each metric $q1$ and $q2$.

each metric, *i.e.*, do there exist individual outliers from the ground truth or are all individual scores a very close to the ground truth score? In Fig 1(a) and 1(b), one observes a difference in the distribution of the computed ratings. In Fig. 1(b), the distribution of ratings is very close to the ground truth rating except for two values. This can be interpreted as a fault in the design of the associated metric q_2 since it fails to accurately predict the actual ratings in two cases. Yet, no such failure is visible in Fig. 1(a). Because of the reduced variance in correlation scores, one may conclude that the IQA algorithm q_1 is globally better designed than q_2.

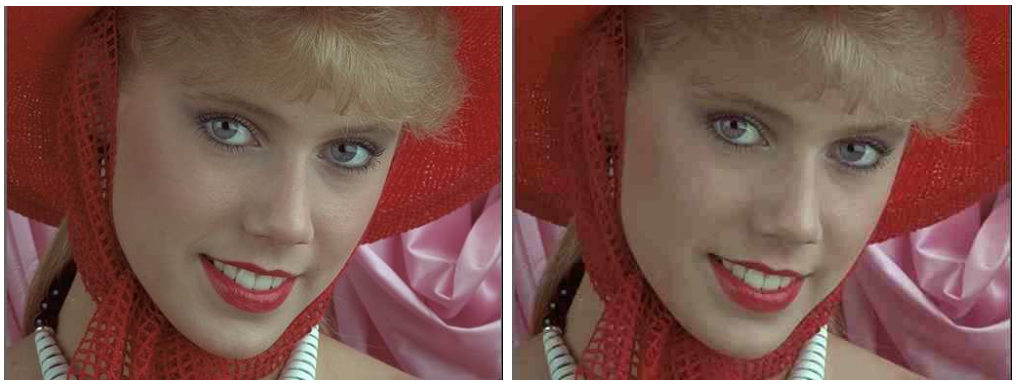
For example, when considering the MS-SSIM index [**?**], one can observe that despite a high degree of correlation with human ratings, it sometimes fails to accurately predict the quality score of a particular image. Fig 2 shows such two cases: 1) both human rating and predicted score of a degraded version of an original image are equivalent and equal to 71, and 2) human rating (54) and predicted score (27) of a degraded version of an original image are different.

Ultimately, however, the interpretation of human ratings is difficult. Suppose, for example, that the human observer rates two compressed (or otherwise distorted) images as 3 and 4 in image quality (on say, a scale of 1 to 10) and also rates two other images as 7 and 8, respectively. Although the difference in rating is the same for both pairs, we have no way to conclude whether the perceived increase in quality between the first pair of images is equal to, greater than, or less than, the perceived increase in quality between the second pair. The subjective ratings only allow us to order the images by quality.

CHARRIER *et al.* [**?**] recently applied a novel psychophysical method, Maximum Likelihood Difference Scaling (MLDS) [**?**], [**?**] that circumvents this limitation of subjective rating methods. MLDS estimates an interval perceptual scale and, thus, makes it possible to quantify supra-threshold perceptual differences between pairs of images in order to evaluate perceptual changes in the images as compression-generated or other distortion is increased. The MLDS method is based on simple, forced-choice judgments and requires remarkably few trials to obtain quantitative estimates of the effects of any degree of distortion [**?**].

In this paper, we evaluate the efficacy of a recently-developed general-purpose IQA algorithm in the specific context of compression-quality trade-off using MLDS. An investigation about its local variation to accurately predict the image quality score is performed, yielding a refinement of the IQA algorithm. The trial IQA algorithm that is used is the MS-SSIM index, due to its high degree of correlation with human ratings [**?**]. This paper is structured as

(a) Both human rating and predicted score of a degraded version (right) of an original image (left) are equivalent and equal to 71



(b) Human rating (54) and predicted score (27) of a degraded version (right) of an original image (left) are different.

Fig. 2. Image extracted from the TID2008 image database for which MS-SSIM is in accordance with human rating (a) and for which MS-SSIM fails to accurately predict the human ratings (b).

follows. In Section II, we present the MLDS method. Section III summarizes MS-SSIM and its relevant parameters. In section IV, we discuss the evolution of the local correlation of the predicted ratings. The apparatus is also presented. In section V, one approach to counterbalance the local lack of correlation is detailed and discussed. Section VI presents the results on two large public image quality assessment databases. This is followed by a concluding section.

II. Maximum Likelihood Difference Scaling

Typical MOS algorithms are based on a psychophysical method introduced by Stevens in 1946 [**?**] known as magnitude estimation. In response to criticisms of the reliability of data collected using magnitude estimation, other scaling methods have been developed, among them the MLDS technique.

The MLDS method is based on forced-choice judgments of stimulus intervals and yields an interval scale of image degradation. The task underlying MLDS is not discrimination of images but direct comparison of suprathreshold differences between pairs of stimuli (images); the observer simply judges which of a pair of stimulus differences is greater. Avoiding the use of rating scales, the MLDS method avoids known problems associated with their use by human beings [**?**], [**?**].

MLDS has previously been used to estimate the effect of distortion level on perceived image quality [**?**]. Next we explain the model of the observer's judgments in the psychophysical task on which MLDS is based, using compression distortion as the application of interest.

An *image series* consists of a *base image* ϕ_1 and compressed versions of the base image denoted $\phi_2, \ldots, \phi_p$, indexed by increasing degree of compression. If image ϕ_i is compressed to a greater degree than image ϕ_j we write $\phi_i > \phi_j$. For brevity, we denote images in the series by their subscripts. The pair (i, j) will serve as shorthand for (ϕ_i, ϕ_j).

On each trial, the observer views two pairs of stimuli (i, j) and (k, l) representing four different levels of compression of the initial image (including possibly no compression). We refer to these two pairs as a *quadruple* denoted $\{i, j; k, l\}$. The observer judges whether the perceptual difference between the first pair (i, j) is greater than that between the second pair (k, l). Over the course of the experiment, the observer judges the differences of a subset of all possible quadruples (pairs of pairs) for the N stimuli in the series $\phi_1, ..., \phi_p$. (i.e., p compression levels).

The goal of MLDS is to assign numerical scale values $(\psi_1, \psi_2, \ldots, \psi_p)$ that can be used to predict how the observer orders the pairs in each quadruple. We refer to these values as a *difference scale*. In principle, we wish to assign these scale values so that the perceived difference between the images of the pair (i, j) is judged greater than the perceived difference between the images of the pair (k, l) if and only if,

$$\|\psi_i - \psi_j\| > \|\psi_k - \psi_l\|. \tag{1}$$

However, if the differences $\|\psi_i - \psi_j\|$ and $\|\psi_k - \psi_l\|$ are close, it is unlikely that human observers would be so reliable in judgment as to satisfy the criterion (1). To take into account this judgment variation, MALONEY and YANG [**?**] proposed a model of difference judgment that allows the observer to exhibit stochastic variations in judgment. We next describe their model. Let $L_{ij} = \|\psi_i - \psi_j\|$ be the *length* of the interval (a_i, a_j). The proposed decision model is an equal-variance, Gaussian, signal detection model [**?**], where the signal is the difference in the lengths of the intervals:

$$\delta(i, j; k, l) = L_{ij} - L_{kl} = \|\psi_i - \psi_j\| - \|\psi_k - \psi_l\| \tag{2}$$

The signal δ is assumed to be contaminated by a Gaussian error ϵ with mean 0 and standard deviation σ to form the judgment variable

$$\Delta(i, j; k, l) = \delta(i, j; k, l) + \epsilon. \tag{3}$$

MALONEY and YANG assumed that the observer, given the quadruple $(i, j; k, l)$, selects the pair (i, j) precisely when $\Delta(i, j; k, l) > 0$. The resulting model of the observer allows for stochastic variation in judgment. When the magnitude of $\delta(i, j; k, l)$ is small relative to the Gaussian standard deviation, σ, the observer, presented with the same stimuli, can give different responses. The degree of inconsistency predicted depends on the magnitude of $\delta(i, j; k, l)$ relative to σ. This dependence can be used to test the model itself [**?**], [**?**].

MALONEY and YANG [**?**] proposed a method to estimate the scale values by direct maximization of the likelihood. However, because the decision rule involves a simple linear combination of the internal responses, the scale values may also be estimated using a Generalized Linear Model (GLM) [**?**], [**?**].

Let R_t be the observers' response to the t^{th} quadruple $(i_t, j_t; k_t, l_t)$ in the experiment, $t = 1, \ldots, n$. R_t is coded as follows: $R_t = 0$ if the difference of the first pair is judged to be larger, and $R_t = 1$ otherwise. The GLM can be specified as

$$g(\mathsf{E}[P(R = 1)]) = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p, \tag{4}$$

where the linear predictor is related to the expected value of the observer's response through a link function, g

$$$**Christophe: Sentence to remove in italic, followed by the suggested one**$$$ *(more details about GLM are given in Appendix A).*

*****BEGIN*******

and X is the model matrix and β a vector of coefficients, as described in Appendix A.

******END********

For binary choice models, as here, the link will be the inverse of a sigmoidal function, and here we use cumulative distribution function (cdf) of a Gaussian.

For each trial, all explanatory variables are set to 0 except for the 4 that correspond to the stimuli presented on that trial. These 4 take the values ± 1 depending on the sign of their contributions to the decision variable, (2). The coefficients, β_i, correspond to the scale values, ψ_i and are estimated by an iterative procedure to yield a maximum likelihood solution.

We estimated difference scales for each observer's data for each image, using MLDS as described above. All computations were carried out in the statistical language R using the `glm` function. We have integrated the functions necessary to perform these fits using either the direct or the GLM approach in an R package (MLDS) available from the Comprehensive R Archive Network (CRAN, accessible from `http://www.rproject.org/`).

If we add a constant c to all the values on the difference scale $(\psi_1, \psi_2, \ldots, \psi_p)$ that maximizes likelihood, the resulting difference scale also maximizes likelihood. If we multiply all the values on the maximum likelihood difference scale $(\psi_1, \psi_2, \ldots, \psi_p)$ by a positive constant $a > 0$, the resulting difference scale also maximizes likelihood once we scale σ by a. Therefore, without loss of generality, we can fix the end points of the maximum likelihood difference scale to be $\psi_1 = 0$ and $\psi_p = 1$. We report all our results in this normalized format.

III. THE TEST IQA ALGORITHM

The MS-SSIM index [**?**] is a multiscale extension of the SSIM IQA algorithm introduced in [**?**]. MS-SSIM contains three factors pertaining to: 1) luminance distortion, 2) contrast distortion and 3) structure comparison.

All of these are first computed within multi-scale subband local patches and then pooled together to obtain the final predicted score between an original image and its degraded version.

The basis of this measure lies in the representation of an image as a vector within an image space. Any image distortion can be interpreted as adding a distortion vector to the reference image vector. In this space, the two

vectors that represent luminance and contrast changes span a plane that is specific to the reference image vector. The image distortion corresponding to a rotation of such a plane by an angle can be interpreted as a structural change.

The luminance comparison is defined as

$$l(I,J) = \frac{2\mu_I\mu_J + C_1}{\mu_I^2 + \mu_J^2 + C_1} \tag{5}$$

where μ_I and μ_J respectively represent the mean intensity of the image I and J, and C_1 is a constant for avoiding instability when $\mu_I^2 + \mu_J^2 \approx 0$. A common choice for the stabilizing constant is $C_1 = (K_1 L)^2$, where L is the theoretical dynamic range of the image's pixels and $K_1 = 0.01$.

The contrast distortion measure is defined to have a similar form:

$$c(I,J) = \frac{2\sigma_I\sigma_J + C_2}{\sigma_I^2 + \sigma_J^2 + C_2} \tag{6}$$

where C_2 is a non negative constant commonly defined as $C_2 = (K_2 L)^2$ ($K_2 = 0.03$), and σ_I (resp. σ_J) represents the standard deviation.

The structure comparison is performed after luminance subtraction and contrast normalization. The structure comparison function is defined as:

$$s(I,J) = \frac{\sigma_{I,J} + C_3}{\sigma_I\sigma_J + C_3} \tag{7}$$

where C_3 is a non negative constant defined as $C_3 = C_2/2$, and $\sigma_{IJ} = \frac{1}{N-1}\sum_{i=1}^{N}(I_i - \mu_i)(J_i - \mu_J)$. Substituting C_3 by $C_2/2$ in 7:

$$s(I,J) = \frac{2\sigma_{I,J} + C_2}{2\sigma_I\sigma_J + C_2} \tag{8}$$

Note that $s(I,J)$ can be negative (e.g., if the subband is inverted)

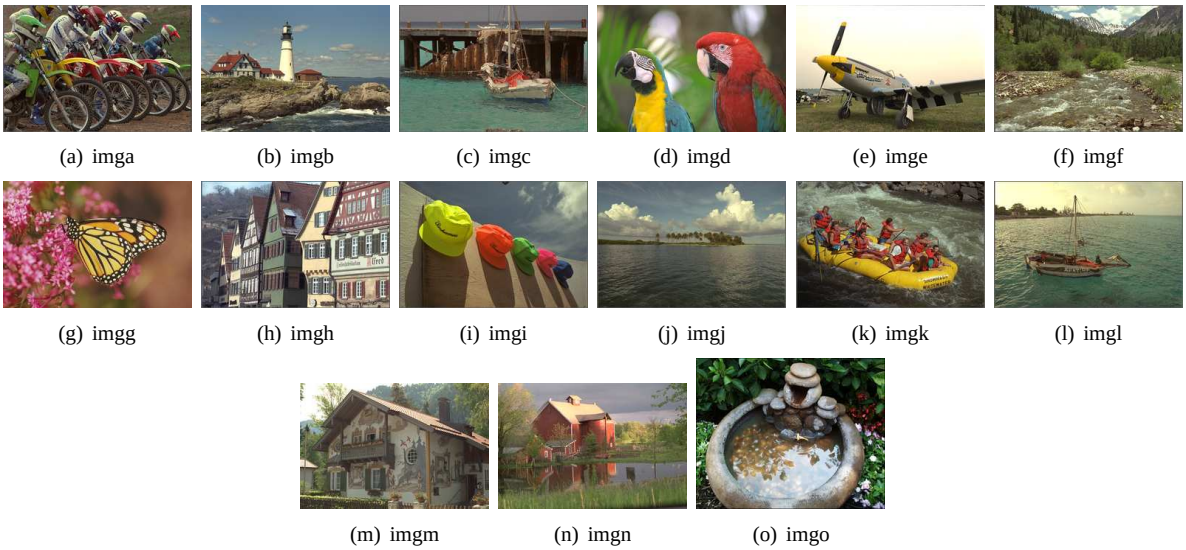
$$$**Christophe: Sentence to remove in italic, followed by the suggested one**$$$ *To obtain a multi-scale index, a low-pass filter is applied to the reference (I) and the distorted images (J). Next a downsampling of the filtered images by a factor of 2 is performed.*

*******BEGIN***** To obtain a multi-scale index, a blur/downsample operation is recursively applied on he reference (I) and the distorted images (J) to generate M scales. *******END*******

The original scale is referred to as scale 1, and the highest scale as scale M. Finally MS-SSIM is given by combining the luminance comparison (5), the contrast distortion measure (6) and the structure comparison (8) at different scales by:

$$\text{MS-SSIM}(I,J) = [l_M(I,J)]^{\alpha_M} \prod_{i=1}^{M} [c_i(I,J))]^{\beta_i} [s_i(I,J))]^{\gamma_i} \tag{9}$$

where the contrast comparison and the structure comparison are computed at the i^{th} scale, and denoted as $c_i(I,J)$ and $s_i(I,J)$, respectively; the luminance comparison $l_M(I,J)$ is computed only at scale *M*. The $2M+1$ exponents α_M, β_i and γ_i, $i = 1, \cdots, M$ are used to adjust the relative importances of the components. In the commonly used



(a) imga (b) imgb (c) imgc (d) imgd (e) imge (f) imgf

(g) imgg (h) imgh (i) imgi (j) imgj (k) imgk (l) imgl

(m) imgm (n) imgn (o) imgo

Fig. 3. The 15 images used in the experiments are shown, with mnemonic labels. For each image, we estimated a difference scale based on each observer's judgments, yielding a total of 450 difference scales.

implementation [**?**], $M = 5$ corresponds to the maximum scale, while $i = 1$ corresponds to the original resolution of the image. In [**?**], the authors defined $\beta_1 = \gamma_1 = 0.0448$, $\beta_2 = \gamma_2 = 0.2856$, $\beta_3 = \gamma_3 = 0.3001$, $\beta_4 = \gamma_4 = 0.2363$, and $\alpha_5 = \beta_5 = \gamma_5 = 0.1333$.

IV. Evolution of the correlation with respect to compression rate

A. Apparatus

Thirty observers participated in the psychophysical tests. All observers had normal color vision (Ishihara test) and normal or corrected-to-normal acuity (Snellen test).

We computed 15 image series using the base images shown in Fig. 3. These images portray a variety of scenes and differ in their distributions of spatial and chromatic detail.

The size of images was typically 768×512 pixels or of similar size. For each visual test, the viewing distance was fixed at 32 pixels per degree of visual angle.

We first tested whether observers could correctly order the compressed images in descending order of quality. If they could not do so, then allowing for possible difficulty in discriminating adjacent images in the scale, there could be no difference scale that could account for their performance.

For an observer to have a valid difference scale, his judgments must satisfy two conditions [**?**], the ordering condition and the six-point condition. If the observer fails either condition, then there is no difference scale that can explain his pattern of choices. For any two stimuli, a_i, a_j we use the notation $a_i \succ_1 a_j$ to mean that the image a_i is judged to be less distorted than the image a_j .

The ordering condition requires only that the observer's ordering of pairs of stimuli must be transitive.

$$(a_i \succ_1 a_j) \ \& \ (a_j \succ_1 a_k) \Rightarrow (a_i \succ_1 a_k).$$

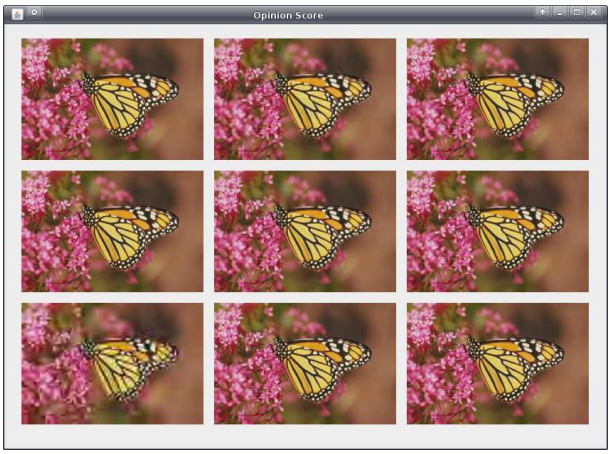


Fig. 4. Example of a single trial during the ordering test. The subject sees an image at the nine trial compression rates. The stimuli are randomly arranged on 3 lines. The subject was asked to order the quality of all images from the best to the worst quality.

for any choice of stimuli a_i, a_j, a_k .Intuitively, a failure of transitive would preclude assigning scale values that predict the observers ordering. We also require that the observers ordering agree with the degree of compression of the stimuli (that the observer judges more compressed stimuli to be more distorted).

The six-point condition is a constraint on how the observer orders differences of pairs of stimuli $(a_i : a_j)$. We use the notation $(a_i : a_j) \succ_2 (a_l : a_m)$ to mean that the observer judges the first pair to be less different than the second. The six-point condition requires that, given any six images $a_i \succ_1 a_j \succ_1 a_k$ and $a_l \succ_1 a_m \succ_1 a_n$, $(a_i : a_j) \succ_2 (a_l : a_m) \& (a_j : a_k) \succ_2 (a_m : a_n)$ implies $(a_i : a_k) \succ_2 (a_l : a_n)$. The condition is effectively a test of additivity of intervals [**?**]. If the observer fails the six-point condition, there is no difference scale that can account for his judgments. Of course, in practice, we must allow for the possibility that observers will make inconsistent judgments due to difficulties in discriminating stimuli. The maximum likelihood fitting methods allow for failures to discriminate. The two conditions are based on the two judgments $\succ_1$, $\succ_2$ and we test the conditions in two experiments reported here.

During this initial test, observers had to first select the highest quality image, then the second highest, etc. A sample trial is shown in Fig. 4. During the test, each time an image was selected by clicking on it, the selected image disappeared and the number of the rank order was shown. If the observer decided to cancel his choice, s/he just had to click on the rank order number. The corresponding image was shown again and the rank order number disappeared. In addition, s/he could deselect more than one image, depending on the selected number, for example, if the observer had already classified six images, the observer could deselect any image numbered from 1 to 6. If s/he deselected image numbered 3, all images from 3 to 6 were automatically deselected.

During the second psychophysical task, the observer saw a quadruple of images drawn from a single image series. These four images were arranged as two pairs (i, j) and (k, l) on a computer display. On half of the trials, the first pair was displayed on the upper half of the display screen, the second on the lower, and on the remaining trials the first pair was displayed on the lower, the second on the upper. For the convenience of the observer,

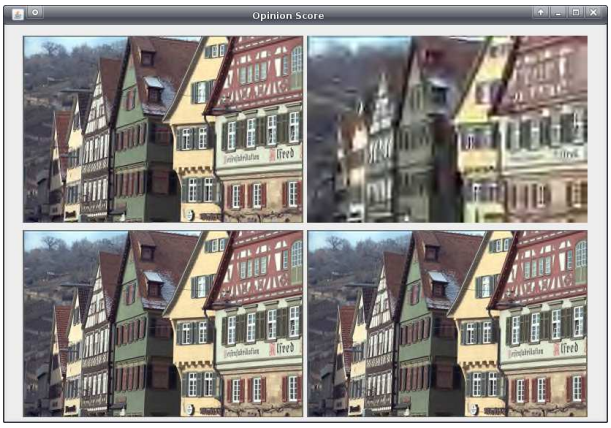


Fig. 5. Example of a single trial in MLDS. The subject was presented with an image at four different compression rates. The stimuli are arranged as two pairs (i, j) and (k, l). In each pair, the right-hand stimulus was more compressed. The subject was asked to judge whether the decrease in quality in going from i to j is greater than the decrease in going from k to l. In this example, most observers would judge that the upper pair exhibits the larger change.

the less compressed of the two images in each pair was always on the left. The observer then judged which pair (upper or lower) exhibited the larger change or difference in quality. A sample trial is shown in Figure 5. Over the course of the experiment, the observer judged several hundred quadruples. These judgments were used to construct a numerical difference scale that captures the effect of additional compression on image quality [**?**], [**?**].

We applied MLDS to evaluate the image quality of the 15 trial original images, each compressed with JPEG2000 to nine different levels: $\{0.1000, 0.3057, 0.5627, 0.7684, 0.9741, 1.1798, 1.3854, 1.5912\}$ bpp, plus the original image. We used the JPEG2000 implementation provided by The JasPer Project [**?**]. We obtained difference scales for each subject and image.

In order to compare MLDS values with scores obtained from the MS-SSIM IQA algorithm, we computed the score provided by the IQA algorithm for each of the nine trial images. Then the difference of scores for each pair of consecutive images was computed. Those differences were then cumulated across the series. The cumulated MS-SSIM scores were then fitted to the MLDS values using a logistic regression function.

B. Results

The obtained results (Fig. 6) show that MS-SSIM captures perceptual changes in images with increasing compression rates very well. Yet, even if MS-SSIM globally yields high correlations with the judgment of human observers, sometimes it fails to accurately predict perceptual changes between images as the compression rate is increased. For example, considering the image `imgk`, observers have judged a high visible difference between stimulus 3 and 4, whereas the associated MS-SSIM values are nearly identical.

In order to investigate these indivudual failures, the same procedure that was used to compare the scores obtained from the IQA algorithm and MLDS values was used for each one of the three factors embedded within MS-SSIM. The results are shown in Fig. 7 for all trial images. The first row of each of the three subfigures corresponds to the contrast comparison values $\prod_{i=1}^{M} c_i(I, J)^{\beta_i}$, the second row corresponds to the luminance comparison values

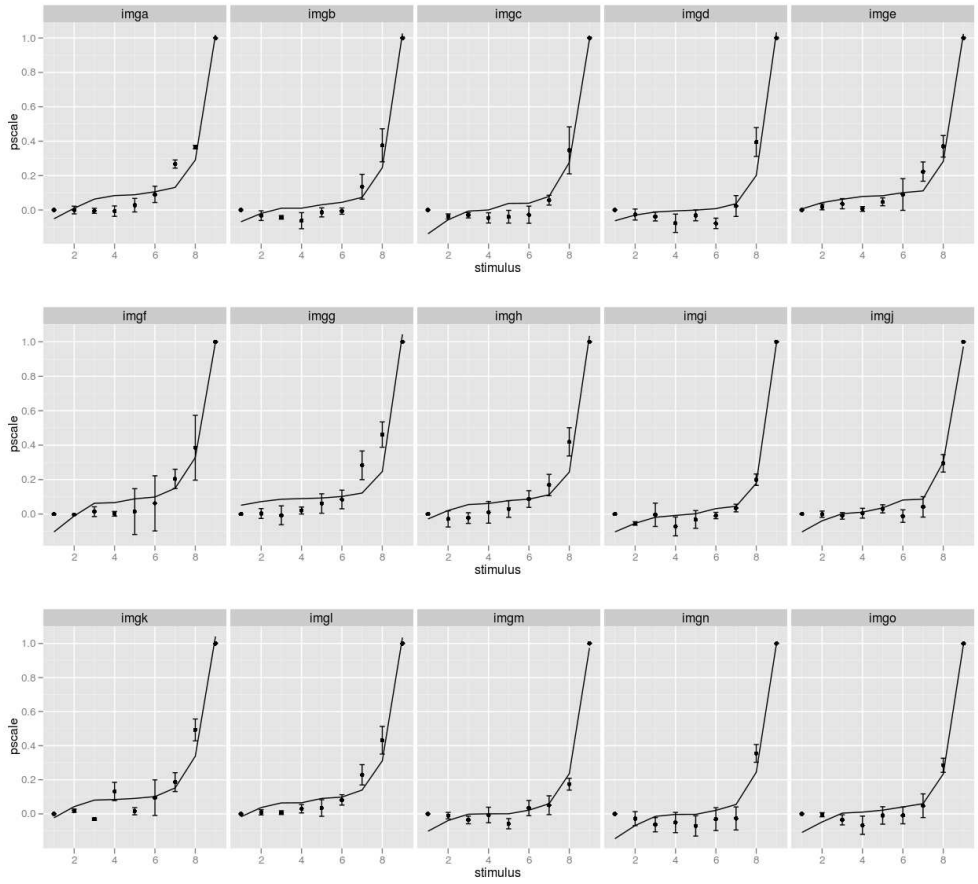


Fig. 6. Obtained results for all trial images and for the thirty observers. The black points and the black curve respectively represent the MLDS and the MS-SSIM values.

$l_M(I,J)^{\alpha_M}$, while the last row represents the structure comparison values $\prod_{i=1}^{M} s_i(I,J)^{\gamma_i}$. At first glance, one might remark that the third factor is less well correlated with MLDS than the two other factors, especially at the beginning of the scale. The same remark can be made when one compares the MLDS values to MS-SSIM in Fig. 6. A poor fit is observed at the beginning of most curves. Thus, structure comparison $\prod_{i=1}^{M} s_i(I,J)$ is of great influence on the MS-SSIM values, as suggested in [**?**].

To acheive the best fit possible, one has to modify the influence of this third parameter. This can be done by changing the five γ_i exponents. To perform this change, we first investigated the influence of the decomposition level M on the fitting with MLDS values.

Since the third (structure) factor is initially computed using $M = 5$ levels, we first investigated the influence of M: how does M influence the curve for this third factor? To measure this influence, we computed the structure comparison factor for levels from 1 to 5. The obtained results are shown in Fig. 8 for a representative subset of trial images (imga to imge), where the black points and the black curve respectively represent the MLDS and the third factor values.

At each decomposition level, one can observe poor fit at the beginning of each scale, for each trial image. This

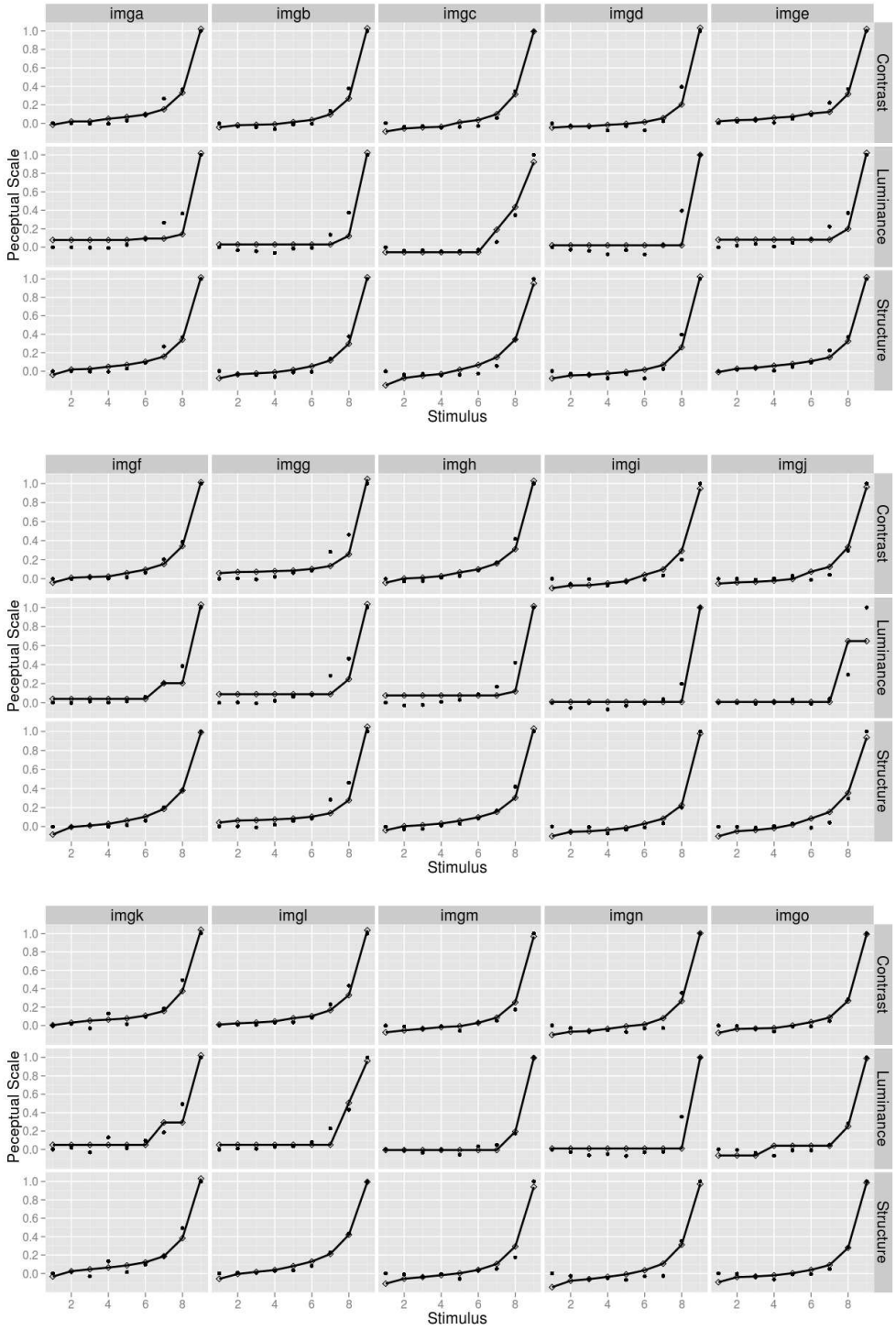


Fig. 7. Obtained results for all trial images. The black points and the curve respectively represent the MLDS and each of the three MS-SSIM factor values. For each of the three subfigures, the first row of each subimage corresponds to the contrast comparison values $\prod_{i=1}^{M} c_i(I,J)^{\beta_i}$, the second row corresponds to luminance comparison values $l_M(I,J)^{\alpha_M}$, and the last row represents the structure comparison values $\prod_{i=1}^{M} s_i(I,J)^{\gamma_i}$

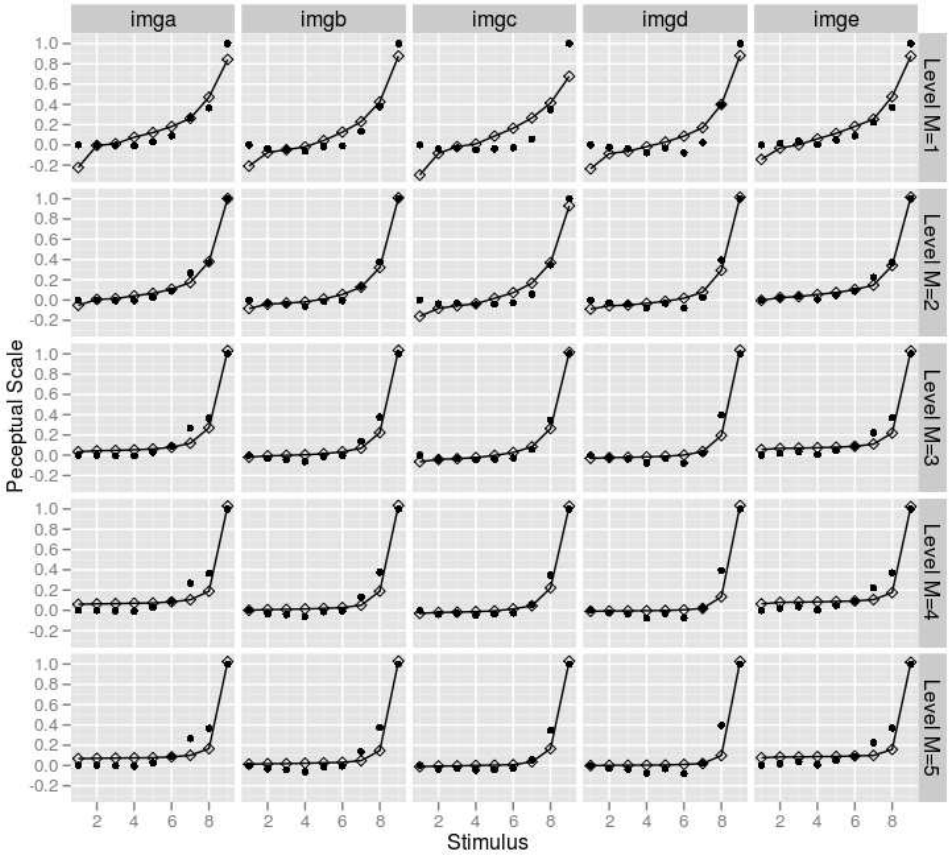


Fig. 8. The structure comparison feature values (8) as used to compute the MS-SSIM values, for different decomposition levels and for the 5 first images (imga to imge). The black points and the black curve, respectively, represent the MLDS and the third factor values used to compute MS-SSIM. Each row corresponds to a decomposition level.

poor fit is observed for low decomposition level values ($M = 1$, $M = 2$). The best fitting curve occurs at the third level, on average.

In order to counterbalance this lack of fit, we first investigated a basic weighting rule that consists of modifying the weight value on the third factor [**?**]. The main goal is to obtain a better fit of the third MS-SSIM values to MLDS. It has been found that refining the exponents values for the third MS-SSIM factor $s(.,.)$, the individual failure observed at the beginning of the scale (Fig. 6) tends to disappear, while the rest of the curve is unaffected, yielding a higher correlation value with human ratings.

From this, it can be presumed that to improve the correlation of the MS-SSIM IQA algorithm scores and MLDS, the coefficients (β_i, γ_i) do not necessarily have to be identical (as initially suggested in [**?**]).

Furthermore, in MS-SSIM luminance is not used at each scale, but only at the coarsest scale, *i.e.*, only at the fifth level with an exponent value equal to $\alpha_5 = \beta_5$. Note that α_5 does not necessarily have to be equal to β_5 and luminance information contained from previous resolution levels could be interesting to take into account to optimize the correlation of MS-SSIM with human ratings. We could take into account all the levels as with the two

other attributes.

Thus, next we investigate the impact of letting all of the parameters, α_i,β_i and γ_i, vary. Thus we will estimate 15 coefficient values to improve the MS-SSIM IQA algorithm.

V. The IQA algorithm genetically improved

Given the obtained results from the weight coefficient κ for the third MS-SSIM factor, we hypothesize that different exponent values for each of the three attributes embedded in the MS-SSIM index would provide a higher global correlation rate.

A. The associated error function

The main objective is to find new exponent values for each decomposition scale of MS-SSIM. The associated formula can be expressed as a 15-parameter function :

$$\text{MS-SSIM}(I, J, \alpha_i, \beta_i, \gamma_i; i = 1, \cdots, M) = \prod_{i=1}^{M} \left[l_i(I, J)^{\alpha_i} c_i(I, J)^{\beta_i} s_i(I, J)^{\gamma_i} \right] \tag{10}$$

where $\sum_{i=1}^{M} \alpha_i + \beta_i + \gamma_i = 1$ and $\forall i \in [1, \cdots, M], 0 \leq \alpha_i \leq 1, 0 \leq \beta_i \leq 1, 0 \leq \gamma_i \leq 1$.

From (10), the search for the new exponent values seeks minimization of the error function

$$E(\alpha_i, \beta_i, \gamma_i; i = 1, \cdots, M) = \min \left(\sum_{j=1}^{K} (\text{MLDS}_j(I, J) - \text{fMS-SSIM}_j(I, J, \alpha_i, \beta_i, \gamma_i))^2 \right) \tag{11}$$

where K is the number of tested images for which the MLDS values are provided, and fMS-SSIM$_j(.)$ are the computed rates obtained following a logistic regression.

In other words, the goal is to estimate the 15 exponent values that minimize the error function $E(.)$. Since the error function is non-convex and may contain numerous local optima, the choice of search strategy to optimize it is important.

B. Search strategy

In this section, the problem of defining a suitable search strategy is addressed. The retrieval of the minimum between the MLDS value and the MS-SSIM value is a global optimization problem, where the error function $E(.)$ is minimized with respect to a set of parameters as in (10). More specifically, the error function (Eq. 11) defines a non-linear multidimensional function, usually characterized by several local maxima. Therefore, the search strategy should find the global minimum, and avoid remaining trapped in local minima. Two problems must be successfully treated 1) the large search space and 2) false matches corresponding to local minima.

The simplest way to find $(\alpha_i, \beta_i, \gamma_i)_{i \in [1, \cdots, M]}$ is by considering a large number of $(\alpha_i, \beta_i, \gamma_i)_{i \in [1, \cdots, M]}$ values, keeping the one whose MS-SSIM value is the closest to MLDS (*i.e.* the one with the lowest error $E(.)$). Of course, the more samples considered, the more precise the end result will be. This kind of brute-force approach based on searching all possible combinations of parameters is not feasible in practice.

The Genetic Algorithm (GA) is a population-based stochastic search procedure that finds exact or approximate solutions to optimization and search problems. Modeled on the mechanisms of evolution and natural genetics, genetic algorithms provide an alternative to traditional optimization techniques by using directed random searches to locate optimal solutions in multimodal landscapes [**?**]. Their basic principles were first introduced by Holland in 1975 [**?**] and extended to functional optimization by De Jong [**?**] and Goldberg [**?**], and have since proven to be efficient and stable in searching for global optimum solutions [**?**], [**?**], [**?**]. One of the most attractive features of GAs is their ability to solve problems involving non-differentiable functions and those defined in discrete as well as continuous spaces.

Usually, a simple GA is composed of three operations: selection, genetic operation, and replacement. GAs use a population, which is composed of a group of chromosomes, to represent the solutions of the system. Defining the solution representation of the system is the first task when applying GAs. The solution in the problem domain can then be encoded into the chromosome in the GA domain, and *vice versa*. Initially, a population is randomly generated. The fitting function then uses values from objective functions to evaluate the quality of fit of each chromosome.

The "fitter" chromosome has the greater chance to survive during the evolution process. The objective function is problem specific; its objective value can represent the system performance index (e.g., an error). Next, a particular group of chromosomes is chosen from the population to be parents. The offspring are then generated from these parents using genetic operations, which normally are crossover and mutation. Similar to their parents, the fitness of the offspring are evaluated and used in replacement processes in order to replace the chromosomes in the current population by the selected off-spring. The GA cycle is then repeated until a desired termination criterion is satisfied, for example, the maximum number of generations is reached, or the objective value is below the threshold.

In this paper, $M = 5$ is the number of levels used to compute the MS-SSIM value. In that case, the GA domain represents a 15-dimensional space in which one point is expressed as $(\alpha_1, \cdots, \alpha_M, \beta_1, \cdots, \beta_M, \gamma_1, \cdots, \gamma_M)$, and the fitness function is defined by (11).

C. Optimization results

To seek each exponent value, the 15 reference images JP2K compressed at nine different compression level as depicted in section IV-A are used to compute the 15 multilevel features $l_i(I, J)$, $c_i(I, J)$ and $s_i(I, J)$ $\forall i \in [1, \cdots, 5]$.

Table I shows the estimated values for each exponent after minimizing (11). Fig. 9 shows the comparison of the MLDS scale values and the 15 parameter fitted MS-SSIM model. The black points represents the MLDS values, the black continuous curves the MS-SSIM indices computed using the original exponent values and the dashed curves by computing MS-SSIM values with the exponent values from Table I. For each trial image (imga to imgo), a better fit to the MLDS values was obtained when the MS-SSIM values are computed with the new exponent values than with the original ones. Table II presents the MSE obtained using the original exponent values and the new ones for all trial images. A reduction of more than 0.2 was attained. This means that the new MS-SSIM indices are better correlated to the MLDS values than the original ones. This is not really surprising, since minimizing

Exponent	α_1	α_2	α_3	α_4	α_5
Value	0.1920	0.2169	0.2026	0.2136	0.1749
CI	[0.0989,0.2415]	[0.1877,0.2791]	[0.1692,0.2384]	[0.1765,0.2868]	[0.0814,0.2304]
Exponent	β_1	β_2	β_3	β_4	β_5
Value	0.9612	0.0097	0.0097	0.0097	0.0097
CI	[0.8288,0.9681]	[-0.0145,0.0933]	[0.0084,0.0112]	[0.0084,0.0112]	[-0.0133,0.1012]
Exponent	γ_1	γ_2	γ_3	γ_4	γ_5
Value	0.0082	0.1586	0.8167	0.0083	0.0082
CI	[0.0073,0.0086]	[0.1241,0.2530]	[0.7250,0.8501]	[0.0073,0.0086]	[0.0073,0.0086]

TABLE I
THE 15 COMPUTED EXPONENTS AND ASSOCIATED CONFIDENCE INTERVALS (CI) WITH A 95% CONFIDENCE LEVEL USING A GA APPROACH UNDER THE CONSTRAINTS $\sum_{i=1}^{M} \alpha_i + \beta_i + \gamma_i = 1$ AND $\forall i \in [1, \cdots, M], 0 \leq \alpha_i \leq 1, 0 \leq \beta_i \leq 1, 0 \leq \gamma_i \leq 1$.

MS-SSIM	Original weighted	New weighted
MSE	0.6092	0.3863

TABLE II
COMPUTED MSE FOR BOTH ORIGINAL MS-SSIM INDEX AND FOR MS-SSIM INDEX USING NEW EXPONENTS USING A LINEAR REGRESSION WITH RESPECT TO MLDS VALUES.

the error function $E(.)$ is the basis for deriving those new values. In addition, confidence intervals with a 95% confidence level are provided for each exponent. They are computed using a boostrap process with 999 replicates.

In order to take into account local correlation to design an IQA algorithm, one can define and minimize an error function between MLDS values and the predicted values. This can help improve the design of the test IQA method, *i.e.* the MS-SSIM measure by finding new exponent values. Fig. 10 displays the original exponent values (black points) and the new ones (black stars) for the three multiscale parameters embedded in MS-SSIM. If we consider the associated coefficients for the structure attribute (third line), we observe that the third decomposition level seems to be considered of greater importance since its exponent value is higher whereas the four others are quite similar. Analyzing Fig. 8, one can see that this level is the best in terms of the fit to the MLDS values. The four other levels are quite similar in terms of fit with the MLDS values. The curve associated with these exponent values is quite similar to the curve associated with the original exponents. Nevertheless, considering individual correlation, the new exponent value associated to the third level is of higher degree than the original one. This is mainly due to the fact that the structure attribute at the third level is a good estimator of the structure degradation evolution.

If we consider the multi-level luminance attribute, only the fifth level was originally considered to be of interest, since only α_5 is used. But, if we observe Fig. 11 the fiting at each level is nearly identical. This suggests that all the five levels should contribute approximately equally to measurement of the luminance degradation.

Focusing on the contrast attribute, note that the first level fits the MLDS values quite well. For levels from 2 to 5, one observes that the values are quite similar with lower fitting accuracy. This agrees with the displayed values in Fig. 10 where the first value is higher than the others and the four last values are identical.

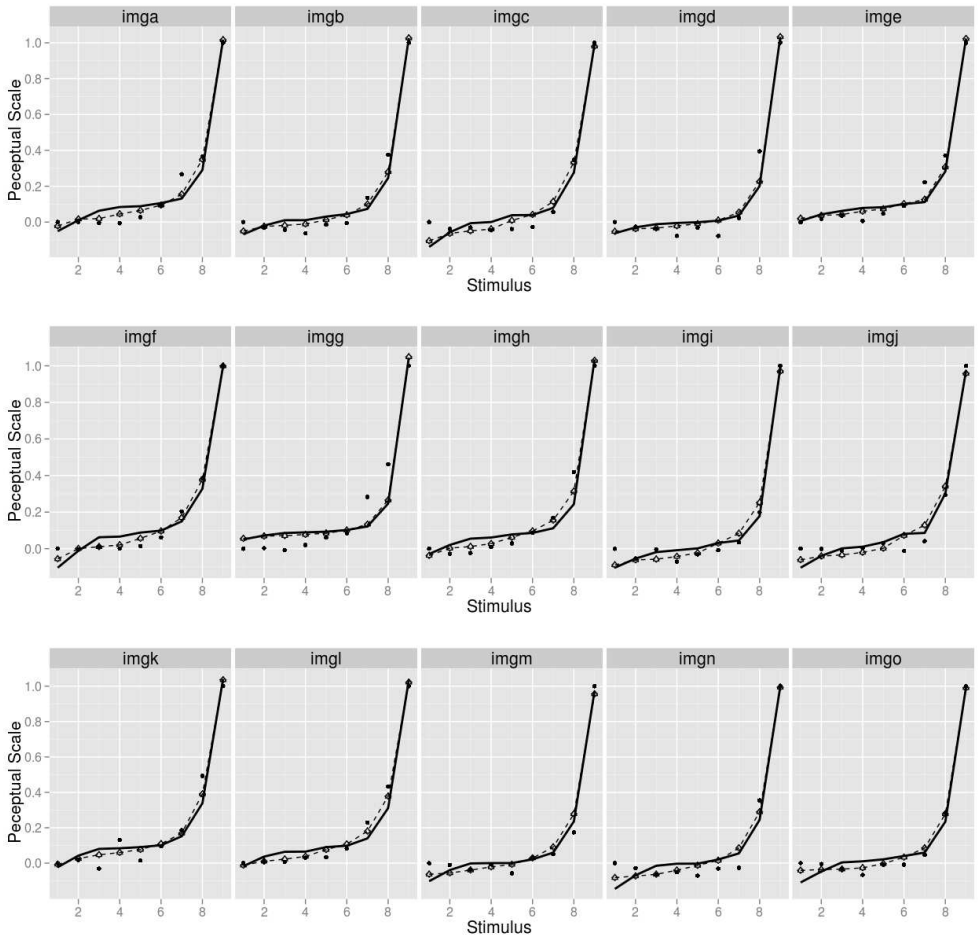


Fig. 9. Obtained results for all trial images. The black points represent the MLDS values, the black curve is associated with the original MS-SSIM index, and the dashed curve represents the computed MS-SSIM values using the new exponent values.

Following the above procedure, better local correlation is obtained, and thus, the error between the MLDS values and the predicted MS-SSIM indices is minimized. This implies that the refined MS-SSIM indices are better correlated to human judgments.

VI. Evaluation of the performance of the refined MS-SSIM index.

In order to judge the relevance of the 15 new exponents estimated in the previous section, we tested the refined MS-SSIM index on both the LIVE and the TID2008 Image Quality databases.

To provide quantitative performance evaluation, three measures of correlation have been used: 1) Pearson, 2) Kendall and 3) Spearman measures. To perform the Pearson correlation measures, a logistic function (as adopted in the video quality experts group (VQEG) Phase I FR-TV test [**?**]) was used to provide a non-linear mapping between the refined MS-SSIM values and subjective scores. We then separately used the subjective scores provided with the overall LIVE and the TID2008 database. Kendall and Spearman correlation measures were computed between the DMOS values and the MS-SSIM indices obtained using both the original exponent values and the new ones

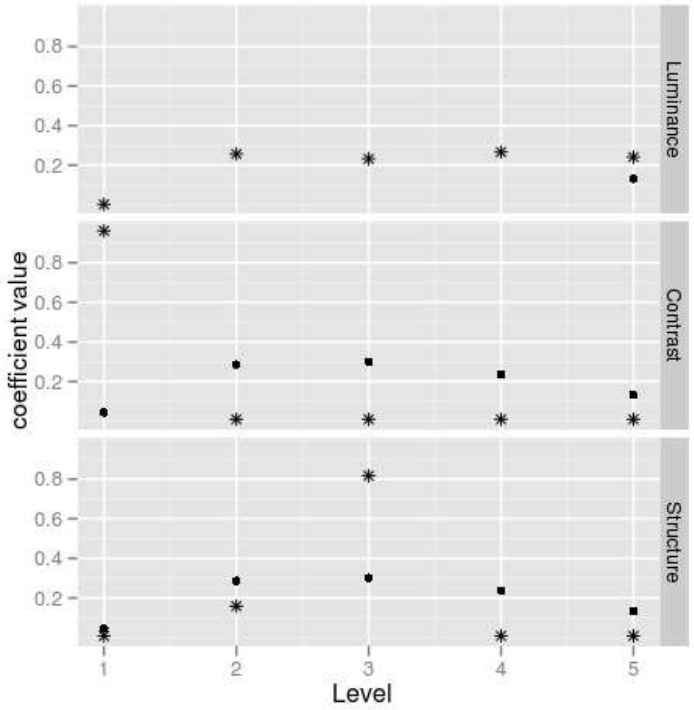


Fig. 10. The original exponent values (black points) and the new ones (black stars) for each MS-SSIM attribute.

(Table I). Those measures can be interpreted as prediction accuracy measures (Pearson and Kendall coefficients) and prediction monotonicity measure (Spearman coefficient).

A. Results on the LIVE database

Considering first the LIVE database, the results are presented in Table III. The scatter plots of DMOS versus both the original and the refined MS-SSIM values are shown in Fig. 13, where each point represents one test image, the vertical and horizontal axes representing MOS and the given distortion objective quality score for the original MS-SSIM (black points) and the refined MS-SSIM values (crosses), respectively.

From both the scatter plots and the correlation evaluation results, we see that the performance of the MS-SSIM index computed with the new exponent values yields improved performance relative to the MS-SSIM values obtained with the original exponent values. This is not true for noisy or blurred images, since a decrease of the correlation coefficients is observed. Nevertheless, when all degradations are included, one observes that the SROCC is significantly higher when new exponent values are used. Naturally, this is driven in part by optimization of QA with respect to JP2K and also FastFading (which uses JP2K), but also JPEG distortion.

B. Results on the TID database

Table IV displays the correlations obtained for both original MS-SSIM index and refined MS-SSIM index with respect to DMOS values from the TID2008 database. When the correlations relative to the subjective values

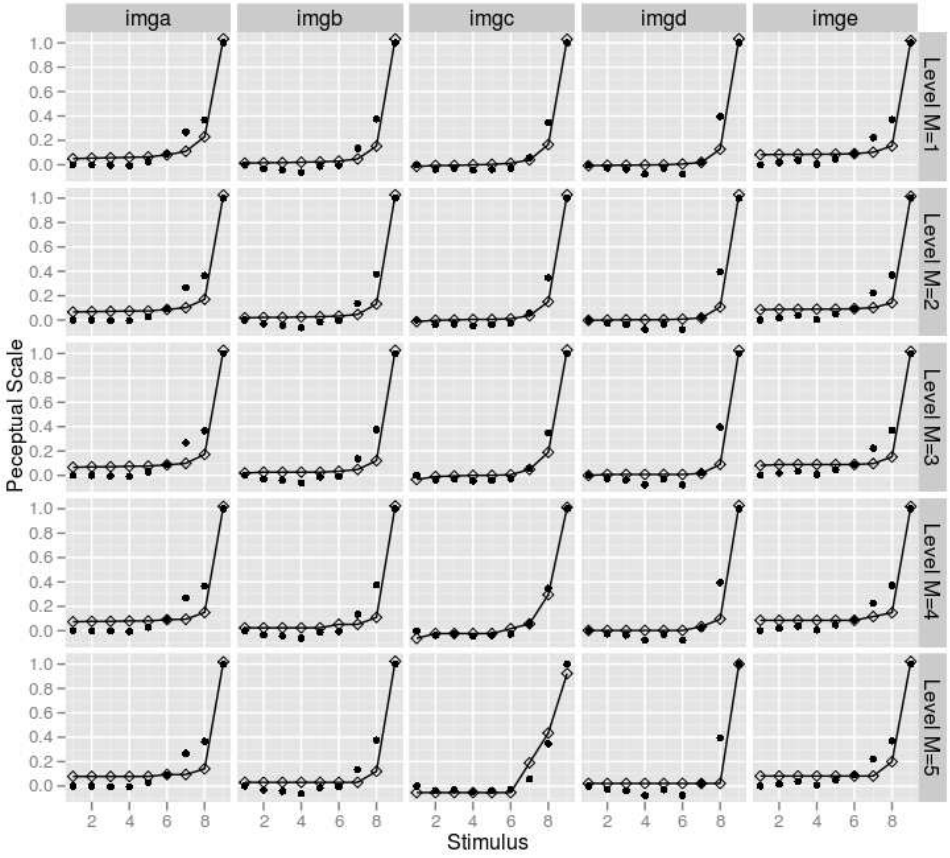


Fig. 11. The luminance feature values for the five decomposition levels and for the five first images (imga to imge). The black points and the black curve respectively represent the MLDS and the luminance factor values used to compute MS-SSIM. The i-*th* row corresponds to a i-*th* decomposition level.

	JP2K		JPEG		White Noise	
	Original	New	Original	New	Original	New
CC	0.783	0.810	0.730	0.742	0.9153	0.9142
KROCC	0.884	0.884	0.849	0.852	0.8887	0.8878
SROCC	0.980	0.991	0.962	0.981	0.9825	0.9813
	Gaussian blur		FastFading		All	
	Original	New	Original	New	Original	New
CC	0.8864	0.8623	0.725	0.788	0.7980	0.8142
KROCC	0.8591	0.8413	0.859	0.876	0.8021	0.8543
SROCC	0.9725	0.9627	0.965	0.974	0.9464	0.9762

TABLE III
COMPUTED CORRELATION COEFFICIENTS FOR BOTH ORIGINAL MS-SSIM INDEX AND FOR MS-SSIM INDEX USING NEW EXPONENTS USING A LINEAR REGRESSION WITH RESPECT TO DMOS VALUES FROM THE LIVE DATABASE.

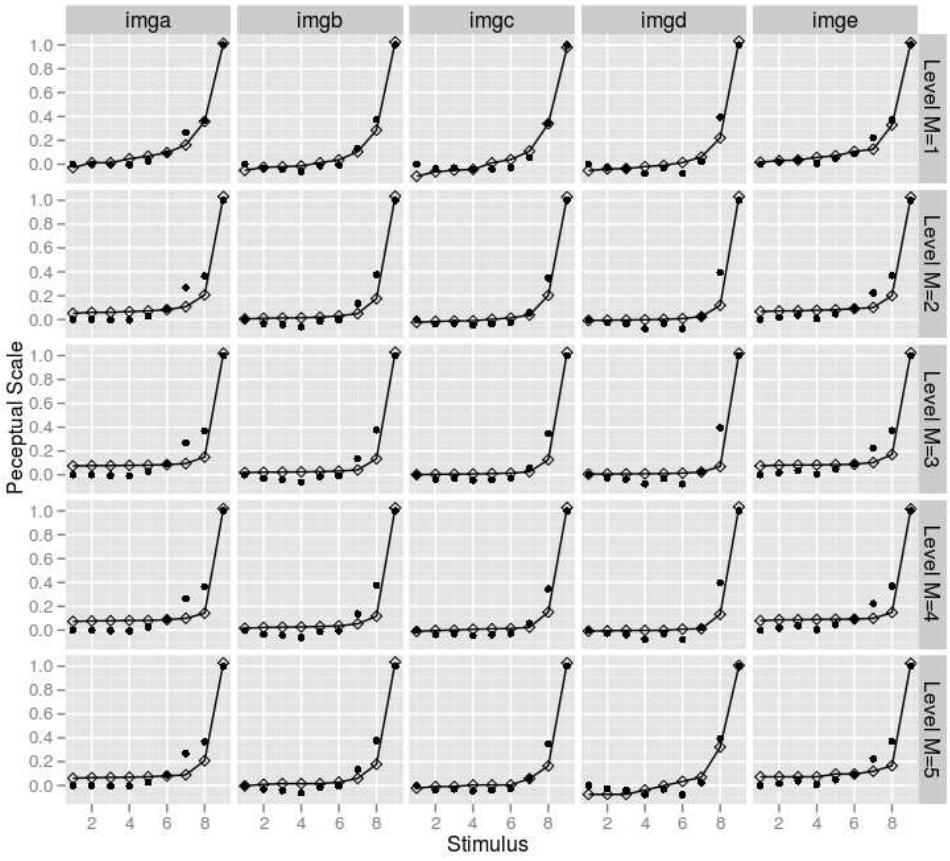


Fig. 12. The contrast feature values for the five decomposition levels and for the 5 first images (imga through imge). The black points and the black curve respectively represent the MLDS and the third factor values used to compute MS-SSIM. The i-*th* row corresponds to a i-*th* decomposition level.

	Degrad #1		Degrad #2		Degrad #3		Degrad #4		Degrad #5		Degrad #6	
	Original	New	Original	New	Original	New	Original	New	Original	New	Original	New
CC	0.7994	0.7700	0.8151	0.7913	0.8278	0.8340	0.8341	0.8224	0.8861	0.8333	0.6672	0.6399
KROCC	0.6139	0.5767	0.6013	0.5677	0.6148	0.6241	0.6117	0.5977	0.6419	0.5887	0.4846	0.4575
SROCC	0.8099	0.7767	0.8055	0.7748	0.8215	0.8265	0.8099	0.7923	0.8706	0.8211	0.6899	0.6547
	Degrad #7		Degrad #8		Degrad #9		Degrad #10		Degrad #11		Degrad #12	
CC	0.8524	0.8355	0.9384	0.9292	0.9638	0.9485	0.9629	0.9796	0.9727	0.9823	0.8784	0.8983
KROCC	0.6569	0.6514	0.8169	0.7793	0.8316	0.8013	0.7489	0.7664	0.8559	0.8876	0.6637	0.6891
SROCC	0.8488	0.8361	0.9563	0.9355	0.9587	0.9458	0.9328	0.9571	0.9697	0.9812	0.8663	0.8852
	Degrad #13		Degrad #14		Degrad #15		Degrad #16		Degrad #17		All	
CC	0.8414	0.8437	0.7417	0.7388	0.7290	0.8666	0.7322	0.7259	0.7721	0.5468	0.8332	0.8532
KROCC	0.6766	0.6957	0.5254	0.5335	0.5038	0.6309	0.5345	0.5427	0.4748	0.4068	0.6577	0.6699
SROCC	0.8609	0.8849	0.7375	0.7434	0.7109	0.8353	0.7239	0.7402	0.6349	0.5430	0.8543	0.8601

TABLE IV
COMPUTED CORRELATION COEFFICIENTS FOR BOTH ORIGINAL MS-SSIM INDEX AND FOR MS-SSIM INDEX USING NEW EXPONENTS USING A LINEAR REGRESSION WITH RESPECT TO DMOS VALUES FROM THE TID2008 DATABASE. THE TYPE OF DEGRADATIONS ARE EXPLAINED IN TABLE V

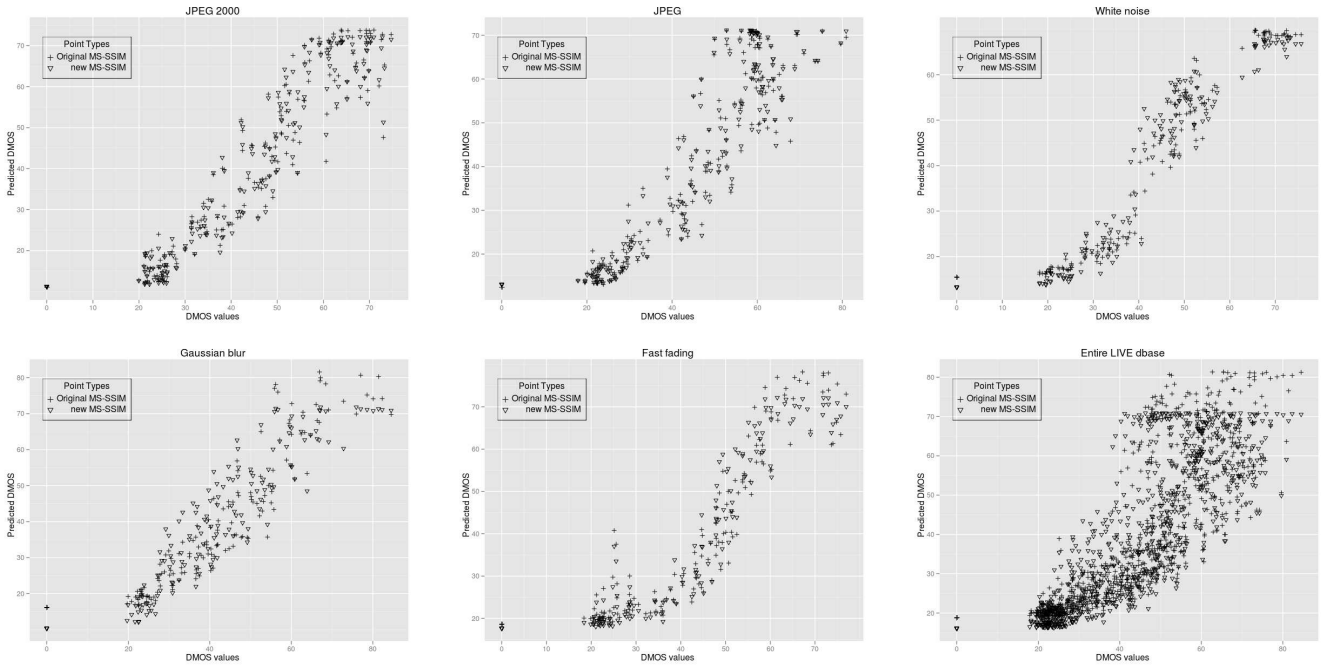


Fig. 13. Scatter plots of DMOS versus the original and MLDS-refined MS-SSIM predictions (with original exponent values and the new ones). Each point represents one test image in the LIVE image database.

calculated on the TID2008 database, the refined MS-SSIM index again outperformed the originally designed MS-SSIM. If we only investigate compression artifacts (degradation #10 to #13, as defined in Table V), the refined MS-SSIM yields a significant increase of the correlation values. Similar to the results obtained from the LIVE image database, a decrease of the correlation values is globally observed for noise artifacts (degradation #1 to #7 and #14). One notes that for two particular noise artifacts, an increase of the correlation value occurs (#3 and #14). A small increase of the correlation values is also obtained for degradation #15 and #16. Both non eccentricity pattern noise (#14) and local block-wise distortions of different intensity (#15) artifacts can be interprated as a block degradation of the image that is a typical compression artifact. This can explain the associated notable increase of the correlation, since the refined MS-SSIM index has been optimized for compression artifacts.

Degradations #3 and #16, respectively, concern a spatially correlated noise and a change of intensity. When analysing the images corresponding to degradation #16, visible differences between the reference image and the degraded versions are not necessarily great. This could correspond to the first part of the obtained curves when the fitting with MLDS values is generated. Actually, from Fig. 9, a flat part is noticeable at the beginning of each curve. The refined MS-SSIM index seems to fit better this particular part than the original MS-SSIM. For JP2K compression artifacts, this particular part corresponds to slightly compressed images, where visible differences are not easily observable (that is the case for degraded images with artifact # 16). This can explain why higher correlation values are obtained for degradation # 16.

Degrad #	Type of distortion
1	Additive Gaussian noise
2	Additive noise in color components is more intensive than additive noise in the luminance component
3	Spatially correlated noise
4	Masked noise
5	High frequency noise
6	Impulse noise
7	Quantization noise
8	Gaussian blur
9	Image denoising
10	JPEG compression
11	JPEG 2000compression
12	JPEG transmission errors
13	JPEG2000 transmission errors
14	Non eccentricity pattern noise
15	Local block-wise distortions of different intensity
16	Mean shift (intensity shift)
17	Contrast change

TABLE V
DESCRIPTION OF THE 17 DEGRADATION TYPES WITHIN THE TID2008 DATABASE

C. Statistical significance

To assess whether the difference in performance between the original MS-SSIM index and the refined MS-SSIM index is statistically significant, we applied a variance-based hypothesis test using the residuals between the DMOS values and the ratings provided by the trial IQA algorithms. This test is based on the F-test that determines whether two population variances are equal. This is done by comparing the ratio of the two computed variances. The null hypothesis is that the residuals from the original MS-SSIM index are statistically indistinguishable (at a 95% confidence level) from the residuals of the refined MS-SSIM. As mentioned in [**?**], the threshold ratio value for which the two sets of residuals are statistically distinguishable can be obtained from the F-distribution [**?**].

The results obtained from this test confirm that the difference of correlation over the entire LIVE database (Table III) is statistically significative.

Regarding the TID2008 database (Table IV), we found that the difference of correlation is not statistically significant overall the database, which is not surprising given the breadth of distortions in the TID database. However, we did find that the refined MS-SSIM index is superior to the original MS-SSIM index with statistical significance for degradations #10 to #13 and #15. Those degradations concern artifacts that occur during a compression scheme applied on images.

VII. CONCLUSION

When one judges the performance of IQA algorithms, correlations with human ratings are computed. The higher the correlation value is, the better the prediction score is. Absolute rating quality methods are usually used to obtain human ratings that will serve as ground truth (MOS or DMOS). Yet, image quality ratings (based on absolute judgments) are considerably less reliable than difference judgements, for reasons described in section II. Intead of

using MOS (or DMOS) values which are obtained from quality ratings based on absolute judgments, we have used a recent psychophysical method, Maximum Likelihood Difference Scaling (MLDS) to evaluate IQA methods and improve them.

We applied it to a large collection of images to assess the consequences of JP2K compression and compared observers' judgments image quality to the predictions of one IQA method, MS-SSIM. We found that MS-SSIM suffers from local failures when assessing JP2K compression, especially due to its third (structure) factor that greatly influences the predicted values. It was found these local failures can be reduced using different values for the three $(\alpha_i, \beta_i, \gamma_i)$ exponents which we estimate from data. The refined MS-SSIM index was found to yield significantly improved performance relative to the original algorithm on two large public image quality assessment databases.

The use of MLDS permits interpretation of the correlation value of IQA algorithms across the series of degradation. This help us identify levels of degradation for which IQA can fail. This is not easily done when absolute rating quality methods are used instead of MLDS. This yields a more precise comparison to human ratings, and helps in the design of high performance IQA algorithms. This allowed us to improve the performance of MS-SSIM for compression-based distortions. Even if the results could be attributed to the use of JP2K compressed images to reweight MS-SSIM, the obtained overall performance for both LIVE and TID2008 database is better than using original MS-SSIM.

APPENDIX A
MAXIMUM LIKELIHOOD DIFFERENCE SCALING AS A GLM

A GLM [**?**] is described by

$$\eta(\mathsf{E}[Y]) = X\beta \tag{12}$$

where $\eta(.)$ is a link function transforming the expected value of the elements of the response vector (Y) to the scale of a linear predictor given by the product of the model matrix (X) and a vector of coefficients (β). The elements Y are distributed as a member of the exponential family. We assume that each quadruple $(i, j; k, l)$ has been reordered so that $i < j < k < l$. Equation (2) can thus be rewritten as

$$\delta(i, j; k, l) = \psi_l - \psi_k - \psi_j - \psi_i \tag{13}$$

The design matrix X can be constructed by considering the weights of the ψ as the covariates. This yields an $n \times p$ matrix X where n is the number of quadruples tested and p is the number of physical levels evaluated over the experiment. On a given trial, the values in only four columns are non-zero, taking on the values 1,-1,-1,1 in that order (these coefficents correpond to the entries in (13)). All the remaining entries are set to 0. For example, consider a set of 7 stimuli distributed along a physical scale and numbered 1 to 7. Four quadruples and the associated

design matrix X are

$$\begin{pmatrix} 1 & 3; & 5 & 7 \\ 7 & 9; & 4 & 5 \\ 1 & 5; & 6 & 7 \\ 2 & 3; & 9 & 10 \end{pmatrix} \begin{pmatrix} 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & -1 & -1 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

To render the model identifiable, however, we drop the first column, which has effect of fixing $\beta_1 = 0$, yielding a model with $p - 1$ parameters to estimate as with the direct method. The GLM can be specified as

$$g(\mathsf{E}[P(R = 1)]) = \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p, \tag{14}$$

where X_j id the j^{th} column of X. In the present case, the responses of the observer can be modeled as Bernouilli random variables. The expected values of the response ,δ, are related to the linear predictors through a non linear function g, that is here the inverse cumulative distribution funtion of the Gaussian. GLM is used to estimate maximum likelihood estimates $\hat{\psi}_2, \cdots, \hat{\psi}_p$, and, together with $\psi_1 = 0$, we have maximum likelihood estimates of the scale values. These form a difference scale where $\sigma = 1$ by assumption, and $\hat{\psi}_p$ is not normalized to 1. As $\hat{\psi}_p = \hat{\sigma}^{-1}$, the scale can be normalized by $\hat{\psi}_p$ as a last step. The justification for these last steps is the invariance of maximum likelihood estimation under reparameterization [**?**]. In practice, the fits were obtained using the open source software R (http://www.r-project.org/) with the package **MLDS** [**?**].